7-1-2007

Localization of ligand binding site in proteins identified in silico

M. Michal Brylinski
Uniwersytet Jagielloński Collegium Medicum

Marek Kochanczyk
Uniwersytet Jagielloński Collegium Medicum

Elzbieta Broniatowska
Uniwersytet Jagielloński Collegium Medicum

Irena Roterman
Uniwersytet Jagielloński Collegium Medicum

Recommended Citation

Brylinski, M., Kochanczyk, M., Broniatowska, E., & Roterman, I. (2007). Localization of ligand binding site in proteins identified in silico. *Journal of Molecular Modeling, 13* (6-7), 665-675. https://doi.org/10.1007/s00894-007-0191-x

J Mol Model (2007) 13:665–675
DOI 10.1007/s00894-007-0191-x

ORIGINAL PAPER

Localization of ligand binding site in proteins identified *in silico*

Michal Brylinski · Marek Kochanczyk · Elzbieta Broniatowska · Irena Roterman

Received: 20 November 2006 / Accepted: 26 February 2007 / Published online: 30 March 2007

Abstract Knowledge-based models for protein folding assume that the early-stage structural form of a polypeptide is determined by the backbone conformation, followed by hydrophobic collapse. Side chain–side chain interactions, mostly of hydrophobic character, lead to the formation of the hydrophobic core, which seems to stabilize the structure of the protein in its natural environment. The *fuzzy-oil-drop* model is employed to represent the idealized hydrophobicity distribution in the protein molecule. Comparing it with the one empirically observed in the protein molecule reveals that they are not in agreement. It is shown in this study that the irregularity of hydrophobic distributions is aim-oriented. The character and strength of these irregularities in the organization of the hydrophobic core point to the specificity of a particular protein's structure/function. When the location of these irregularities is determined versus the idealized *fuzzy-oil-drop*, function-related areas in the protein molecule can be identified. The presented model can also be used to identify ways in which protein–protein complexes can possibly be created. Active sites can be predicted for any protein structure according to the presented model with the free prediction server at http://www.bioinformatics.cm-uj.krakow.pl/activesite. The implication based on the model presented in this work suggests the necessity of active presence of ligand during the protein folding process simulation.

Keywords Hydrophobic collapse · Protein folding · Active site · Ligand binding

M. Brylinski · M. Kochanczyk · E. Broniatowska · I. Roterman
Department of Bioinformatics and Telemedicine,
Jagiellonian University – Collegium Medicum,
Łazarza 16,
31-530 Krakow, Poland

M. Brylinski
Faculty of Chemistry, Jagiellonian University,
Ingardena 3,
30-060 Krakow, Poland

M. Kochanczyk · I. Roterman (✉)
Faculty of Physics, Astronomy and Applied Computer Science,
Jagiellonian University,
Reymonta 4,
30-059 Krakow, Poland
e-mail: myroterm@cyf-kr.edu.pl

Introduction

Since the classic work by Kauzmann [1], hydrophobic interactions have been confirmed to play a crucial role in forming and stabilizing the protein tertiary structure [2–5]. It is generally accepted that globular proteins consist of a hydrophobic core and a hydrophilic surface [6, 7]. The way in which the amino acid sequence partitions a protein into its interior and exterior has been described [8]. The inside regions are densely packed chain sites where hydrophobicity is observed to be at a local maximum, whereas the outside regions correspond to less densely packed sites where hydrophobicity is at a local minimum. The spatial distribution of amino acid hydrophobicity has been used to differentiate native and non-native protein structures [9–12]. Irbäck and co-workers brought forward a proof for nonrandom hydrophobicity structures in protein chains [13]. A second-order hydrophobic moment was discussed for description of protein hydrophobicity [14]. Detailed analyses of the spatial variation of hydrophobicity, focused on the region of transition between the protein interior and exterior, were carried out for 30 relatively diverse globular

proteins as well as for 14 decoys [15]. Apart from soluble proteins, the distribution of apolar and polar residues has provided comprehensive information about transmembrane protein architecture [16–19]. The hydrophobic effect has been suggested to be the dominant driving force in protein folding [20–22].

The model presented in this paper is oriented on localization of the area responsible for ligand binding or protein–protein complex creation, based on the characteristics of the spatial distribution of hydrophobicity in a protein molecule. It is assumed that hydrophobicity changes from the protein interior (maximal hydrophobicity) to the exterior (close to zero hydrophobicity) according to the three-dimensional Gauss distribution. It is generally accepted that the core region is not well described by a spheroid of buried residues surrounded by surface residues due to hydrophobic channels that permeate the molecule [23, 24]. This being so, we should be able to identify regions with high deviation versus the ideal model by making a simple comparison of the theoretical (idealized according to the Gauss function) and empirical spatial distribution of hydrophobicity in a protein. The regions recognized by high hydrophobicity density differences seem to reveal functionally important sites in proteins.

The model presented here can be used for structure-based prediction of the localization of active sites in proteins of unknown function. It can also be used for qualitative and quantitative analysis of known protein–protein or protein–ligand complexes. The statistical method introduced for assignment of theoretical hydrophobicity to small ligands, combined with searches for equivalent hydrophobic cavities, may also be very useful for docking simulations.

Materials and methods

Data

The following structures were selected for analysis of proteins complexed with small ligands: myoglobin (PDB ID: 1A6M) [25], subtilisin DY (PDB ID: 1BH6) [26], carboxypeptidase A2 (PDB ID: 1DTD) [27], chymotrypsin (PDB ID: 1GG6) [28], c-type lysozyme (PDB ID:1LMQ) [29] and ribonuclease (PDB ID: 1RGE) [30]. The following CAPRI targets were selected for studying the structures of protein–protein complexes: transcriptional antiterminator protein LicT (target 09, homodimer, PDB ID: 1H99) [31], cohesin-dockerin complex (targets 11 and 12, PDB ID: 1OHZ) [32] and the complex between protein serine/threonine phosphatase-1 (delta) and N-terminal domain of the myosin phosphatase targeting subunit MYPT1 (target 14, PDB ID: 1S70) [33].

Theoretical (expected) hydrophobic core

The *fuzzy-oil-drop* model, which was applied to simulate so-called hydrophobic collapse in the protein folding process, represents the theoretical, idealized hydrophobic core of the protein molecule. Such a hydrophobic core surrounded by the spheres of gradually decreased hydrophobicity is assumed to be represented by a three-dimensional Gauss function:

$$Ht_j = \frac{1}{Ht_{sum}} \exp\left(\frac{-(x_j-\bar{x})^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_j-\bar{y})^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_j-\bar{z})^2}{2\sigma_z^2}\right) \quad (1)$$

The value of Ht_{sum} represents the sum of theoretical hydrophobicity of all the grid points. The value of the probability distribution (as the value of the Gauss function is usually interpreted) Ht_j is assumed to represent the hydrophobicity density for the j-th grid point in the *fuzzy-oil-drop*. The hydrophobicity maximum is localized in the center of the ellipsoid and decreases in a distance-dependent manner according to the three-dimensional Gauss function. The mean value at which the Gauss function reaches its maximum is localized at the (0,0,0) point in a coordinate system. The standard deviation represents the size of the drop (according the three-sigma rule) depending on the length of the polypeptide under consideration.

The protein molecule is localized with its geometrical center at the origin of the coordinate system. The longest distance between effective atoms (side chains represented by the geometrical centers of the atoms present there) determines the Z-axis. The Y-axis is oriented according to the longest distance between the projections of the effective atoms on the XY plane. The longest distance between the projections of the effective atoms along the X-axis determines the size of the drop.

For the orientation described above, the σ_x,σ_y,σ_z values can be calculated. The value of the longest distances versus the (0,0,0) point along each axis increased by 9 Å (cut off distance for hydrophobic interaction) divided by 3 gives value of appropriate σ.

A grid (its size depends on the molecule size) is created in three-dimensional space. The values of the Gauss function can be calculated for each grid point. The values of the Gauss function (representing the hydrophobicity distribution) are standardized to give a value of the sum of all values over all grid points equal to 1.0. The system of grid points calculated in this way is treated as the ideal *fuzzy-oil-drop* with a hydrophobicity distribution according to the Gauss function.

The Gaussian function makes it possible to calculate hydrophobicity in each point of space. The grid system can be created in step-wise form. In particular, the positions of effective atoms can also be treated as grid points, which collect the hydrophobic interaction of particular residue (*j*-th) with all others localized closer than the cutoff distance, and can further be treated as parameters describing a particular amino acid (*j*-th). This is why the distribution of hydrophobicity density can be presented in the form of a profile along the polypeptide.

Empirical oil-drop observed in proteins

The same grid points as described above are used to calculate the empirical hydrophobicity distribution representing the empirical oil drop. The empirical hydrophobicity attributed to each grid point represents interaction with all effective atoms (below the cutoff distance). Generally, the grid point by itself represents zero hydrophobicity. If the grid point is localized in the place of effective atom of particular residue, its hydrophobicity is equal to the hydrophobicity of this residue. The observed hydrophobicity is calculated according to the simple sigmoid function previously proposed to quantitatively describe hydrophobic interactions [34]. The *j*-th point collects hydrophobicity Ho_j as follows:

$$Ho_j = \begin{cases} \frac{1}{Ho_{sum}} \sum_{i=1}^{N} \left(H_i^r + H_j^\gamma\right)\left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right], & \text{for } r_{ij} \le c \\ 0, & \text{for } r_{ij} > c \end{cases} \quad (2)$$

where Ho_j represents the empirical hydrophobicity value characteristic for the *j*-th grid point, H_i^r represents the hydrophobicity characteristic of the *i*-th amino acid, r_{ij} is the distance between the *j*-th effective atom (generally *j*-th grid point) and *i*-th effective atom of the amino acid, and *c* expresses the cutoff distance, which has a fixed value of 9.0 Å following the original paper [34]. The value of Ho_{sum} represents the sum of observed hydrophobicity of all the grid points.

Applying this function requires attribution of hydrophobicity parameters to each amino acid. Many scales for residue hydrophobicity are available. Some of them are based on analysis of known protein 3D structures [6, 14, 35–38], while others are derived from the physicochemical properties of amino acid side chains [39, 40]. Selection of an appropriate scale seems crucial, so a new statistics-based hydrophobicity scale for amino acids has been created.

The differences between idealized and empirical oil-drop

Since the theoretical *Ht* and observed *Ho* distributions of hydrophobicity are both standardized to 1.0, these two distributions can be compared. The differences between the theoretical and empirical distributions ΔH_i express the irregularity of hydrophobic core construction. For the *i*-th residue, ΔH_i is calculated as follows:

$$\Delta H_i = Ht_i - Ho_i \quad (3)$$

where Ht_i and Ho_i are the theoretical and observed values of hydrophobicity for the geometric center of the *i*-th residue, respectively. The theoretical *fuzzy-oil-drop* and empirical oil drop were calculated for all proteins taken into consideration. A color scale was introduced to express the magnitude of the difference ΔH_i in a particular protein area, visualizing the localization of these discrepancies in the protein molecule. The one-dimensional profiles of ΔH_i were smoothed by averaging of the raw data using a five-residue running window frame. This simple method gives a legible curve without affecting the positions of dominant local extrema.

Comparative analysis

Method oriented on active site recognition SuMo [41] (http://www.sumo-pbil.ibcp.fr/cgi-bin/sumo-welcome) was applied for comparative analysis. This method is based on the comparisons against binding sites from the PDB as criterion for active site recognition. The SuMo method is a two-step procedure. The representation of a protein structure by a set of chemical groups (unbound hydrogen bond donors or acceptors, accessible sides of aromatic rings and carboxylate, primary amide, etc.) is performed as a first step. The comparison of preformatted data is made in the second step. As a result, the list of similar active sites ordered according to decreasing score corresponding to the size of the matched sites is formed.

The results of SuMo and *fuzzy-oil-drop*-based were compared with the protein crystal structure. The distances below 12 Å between the ligand centre of gravity and Cα of sequential residues were used to define the amino acids responsible for ligand binding.

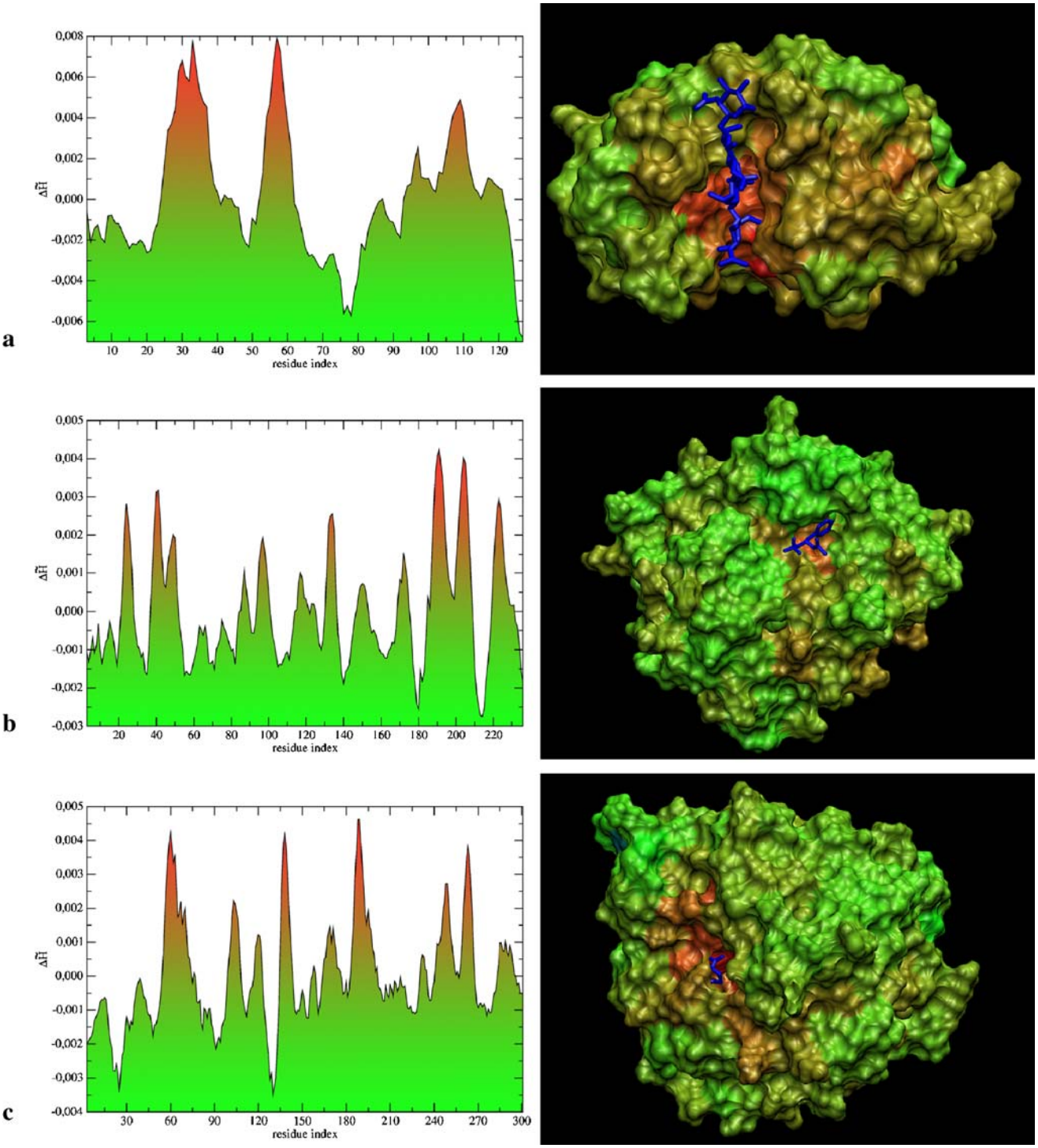

Fig. 1 One-dimensional profiles of ΔH per amino acid (*left column*) and three-dimensional distribution of ΔH on protein surface (*right column*) for lysozyme complexed with N-acetylglucosamine (**a**), chymotrypsin complexed with N-acetyl-L-phenylalanyl trifluoromethyl ketone (**b**), carboxypeptidase A2 inhibited by leech carboxypeptidase inhibitor (**c**), subtilisin DY inhibited by N-benzyloxycarbonyl-Ala-Pro-Phe-chloromethyl ketone (**d**), myoglobin complexed with heme (**e**) and ribonuclease complexed with guanosine-2′-monophosphate (**f**). The *color scale* expresses the magnitude of difference in a particular protein surface area. The dark blue (*thick line*) ligands are localized at their binding sites according to crystal structure. The 3-D structures were received using the VMD program with ΔH_i values put into the β-factor column and taken as criteria for the *color scale*

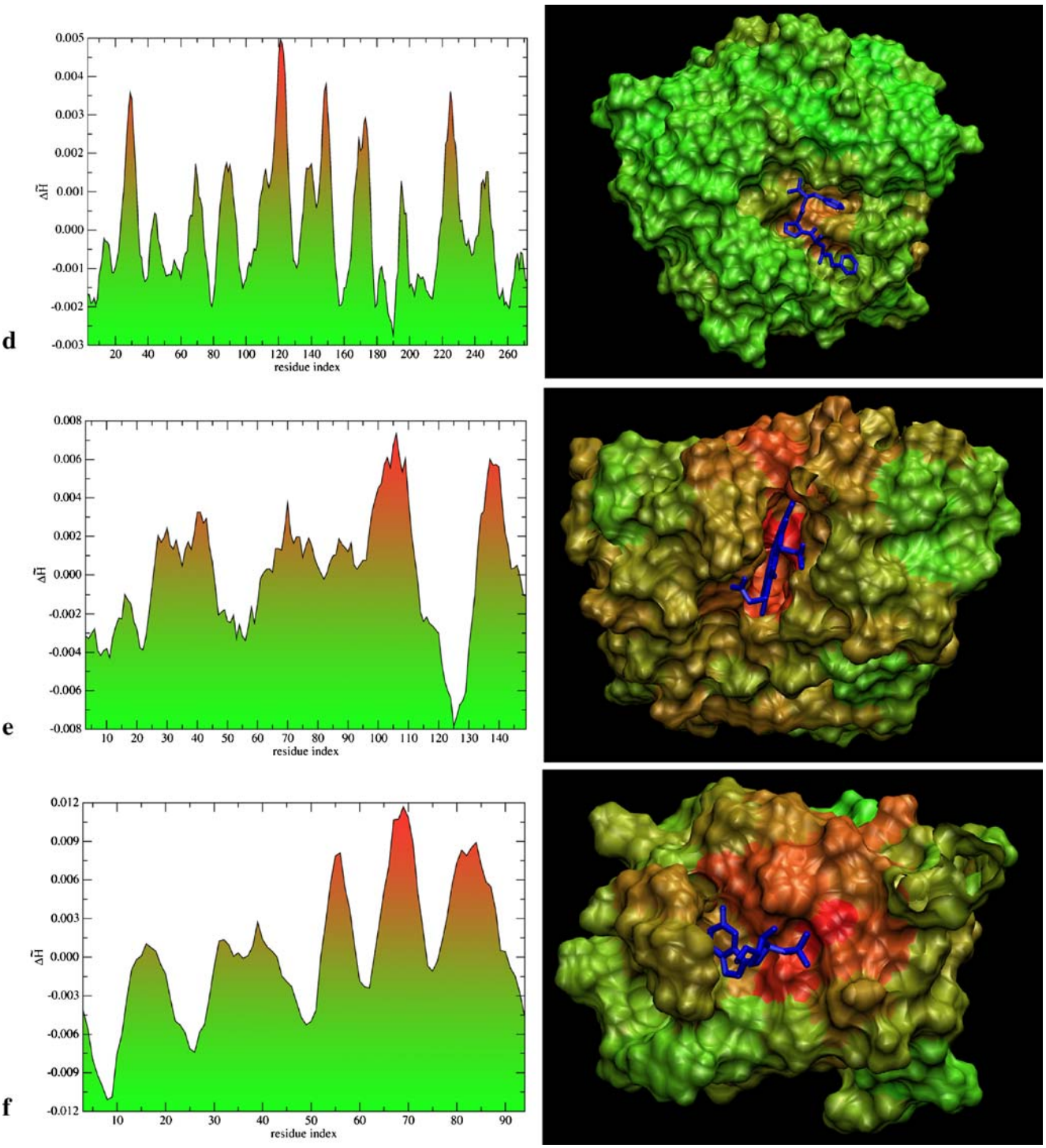

Fig. 1 (continued)

Results

Comparing the expected and observed oil drops

The results of comparing the idealized and empirical hydrophobicity distributions are shown in two forms: a one-dimensional profile of ΔH_i (as dependent on the localization of the amino acid in the polypeptide chain), expressing the magnitude of differences between expected and observed hydrophobicity; and the three-dimensional distribution of ΔH_i (as it appeared on a particular part of the protein surface). Figure 1 shows the results describing six proteins

selected for analysis of proteins complexed with small ligands. The left column represents the one-dimensional profiles of ΔH_i. The color scale also visualizes these profiles. The same color scale adopted for the three-dimensional distribution of ΔH_i is shown in the right column. The dark blue (thick line) ligands, localized at their binding sites according to crystal structure, reveal very good agreement between the observed and predicted areas for potential interaction sites.

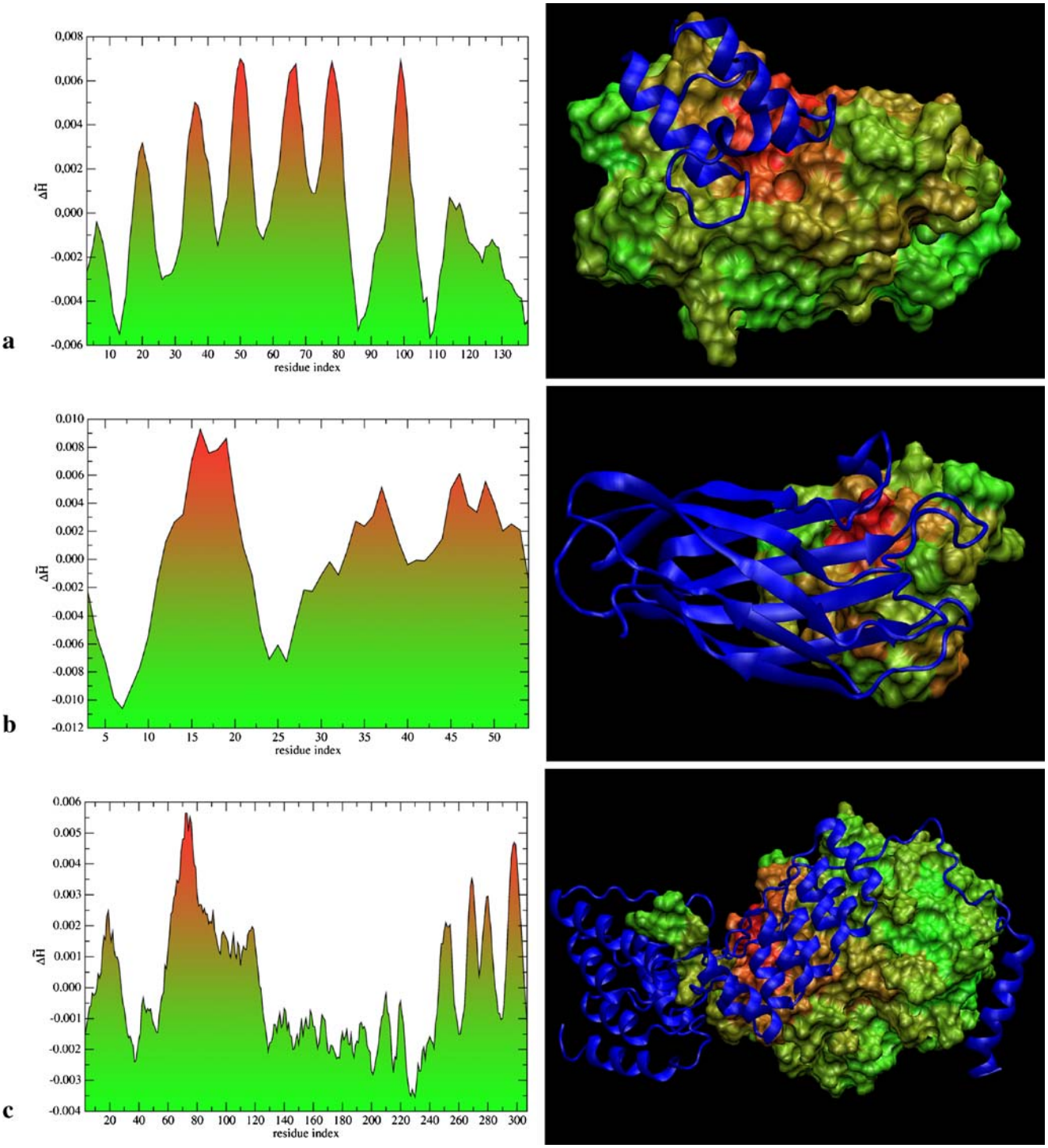

Fig. 2 One-dimensional profiles of ΔH per amino acid (*left column*) and three-dimensional distribution of ΔH on protein surface (*right column*) for cohesin-dockerin complex: results for cohesin (**a**), results for dockerin (**b**); complex between protein Ser/Thr phosphatase 1 and MYPT1: results for Ser/Thr phosphatase (**c**), results for MYPT1 (**d**); and LicT homodimer (**e**). The *color scale* expresses the magnitude of difference in a particular protein surface area

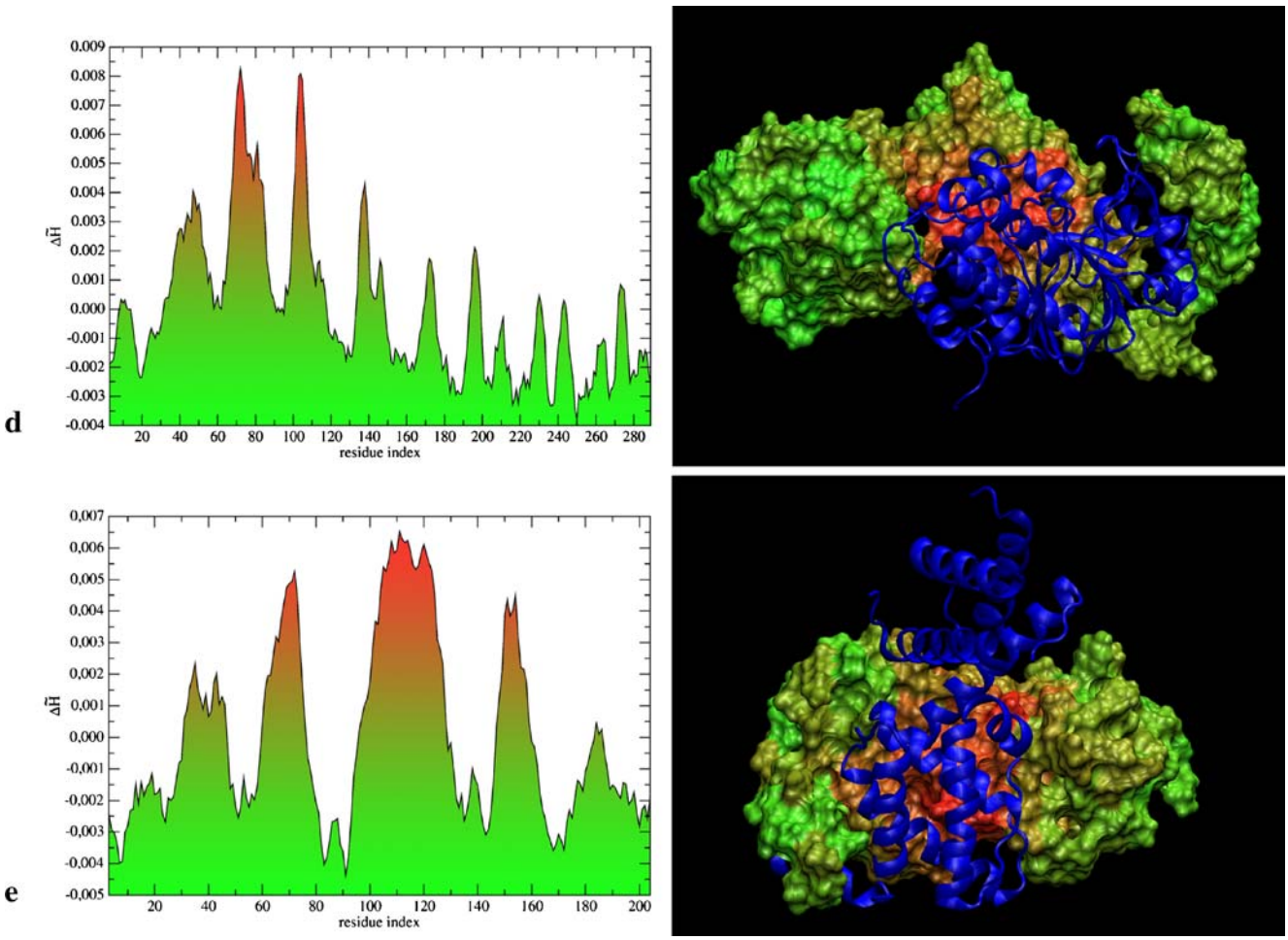

Fig. 2 (continued)

Localization of biological function

Figure 2 also shows the localization of the area critical for the biological function of protein molecules. The proteins represent the case which can be called protein–protein complex construction. The profile of ΔH_i and its three-dimensional distribution as it appears in the protein complex are shown. Both partners' binding sites can be easily recognized. The irregularity of the hydrophobicity distribution for the surface (hydrophobicity higher than the expected), and the cavity (hydrophobicity lower then expected) disturbing the regularity of the hydrophobicity distribution, seem to be good markers for mutual interaction.

Comparative analysis

The amino acids recognized as active site by SuMo and *fuzzy-oil-drop* model are given in Table 1. The residues identified on the basis of distance criterion are also given in Table 1.

The protein–protein complexes shown in the Fig. 2 have not been taken for this analysis due to SuMo limitation, which excludes the protein–protein complex creation.

The residues distinguished by *fuzzy-oil-drop* criterion are directly identified by maxima shown in Fig. 1. Two options are possible for the SuMo model: with and without ligand defined. To make the comparison complete, both forms were applied in this analysis. The score values are given in parenthesis. The residues identified on the basis of the crystal structure with distance (between Cα position and ligand center of gravity). Figure 3 illustrates the comparison of the residues identified as belonging to active site.

The results of comparative analysis seem to represent rather high accordance particularly taking into account the character of the criteria used in these two methods. They seem to be quite differentiated: the protein molecule-oriented in *fuzzy-oil-drop* and ligand-oriented in SuMo.

Table 1 Numbers of residues recognized as belonging to active site according to: the close vicinity of residues versus the ligand center of gravity defined as distance below 10 Å and below 15 Å (numbers in parenthesis), according to SuMo in form of ligand defined and ligand not defined (the score values are given in parentheses) and according to *fuzzy-oil-drop*

Protein	Crystal-based D<10 Å (D<15 Å)	SuMo		*Fuzzy-oil-drop*
		Ligand defined	Ligand undefined	
1A6M	42, 43, 64–72, 89–93, 97–99, 103, 104, 107 (25, 28–33, 39–46, 61–72, 86–108, 139, 142)	HEM: 92, 93, 97 (3.050) HEM: 92, 93, 97 (3.050) HEM: 42, 43, 44, 45 (2.700)	HEM: 92, 93, 97 (3.050) HEM: 92, 93, 97 (3.050) FMN: 39, 40, 41, 103 (2.727)	39–44, 68–71, 96–110, 134–141
1RGE	35–41, 54, 85, 86 (31–44, 49–57, 60–72, 80–89)	2GP: 36–41, 54, 65, 69, 85, 86 (11.013) 2GP: 36–41, 54, 65, 69, 85, 86 (10.188) 2GP: 65, 69, 86 (3.250)	2GP: 36–41, 54, 65, 69, 85, 86 (11.013) 2GP: 36–41, 54, 65, 69, 85, 86 (10.188) SGP: 37–41, 65, 69, 86 (9.296)	54–57, 65–72, 79–87
1LMQ	35, 44–46, 50–52, 57–61, 107–110 (31–36, 42–64, 69, 72, 80, 95, 98, 99, 103–114)	NAG: 3, 85, 86 (2.450) NAG: 40, 83–85 (2.400) NAG: 40, 83, 85 (2.400)	SUC: 35, 52, 57, 59, 107–109, 112 (6.738) BUL: 35, 52, 57, 59, 107–109, 111 (6.738) NAG-NAG: 35, 52, 57, 59, 107–109, 111 (6.286)	26–37, 54–60, 105–111

Discussion and conclusion

Many proteins of unknown biological function, identified on the basis of genome analysis, await a unified automated method for determining their biological activity [42]. The next step is to develop methods able to predict a protein's function from an examination of its structure. Some of the techniques used to identify functionally important residues from the sequence or structure are based on searching for homologues of proteins of known function [43–46]. However, homologues need not have related activity, particularly when the sequence identity is below 25%

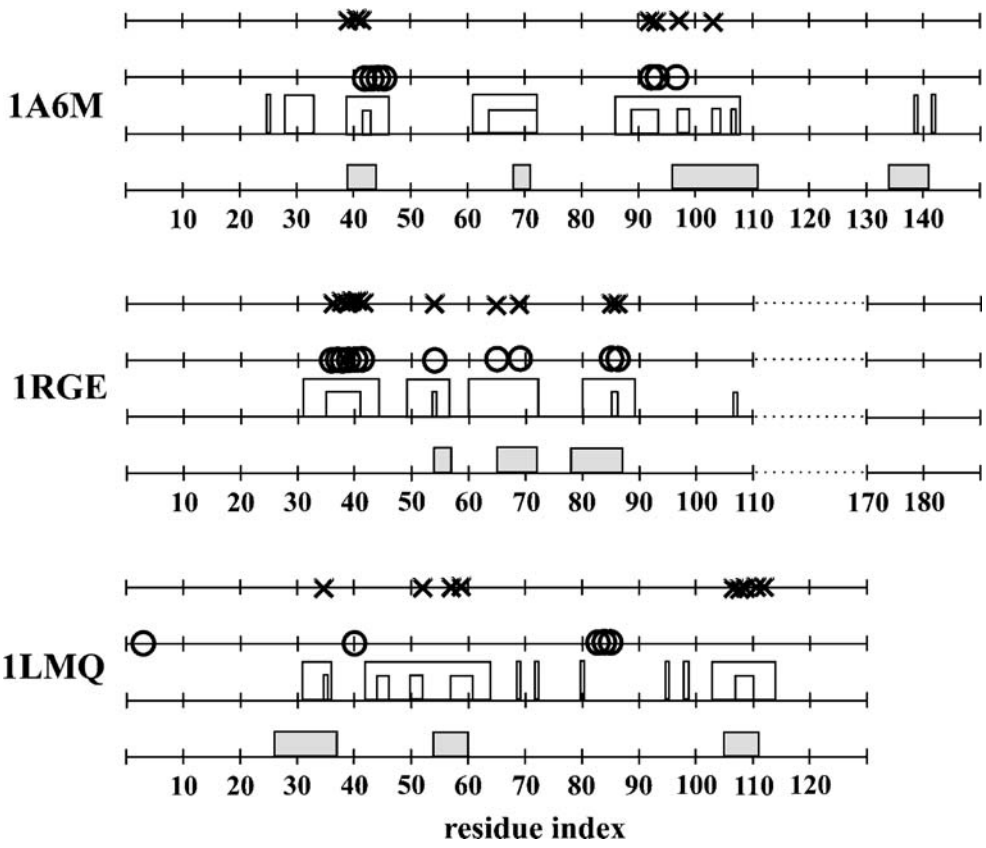

Fig. 3 Comparison of the residues recognized as active site by (bottom to top): *fuzzy-oil-drop* method, crystal structure for the distance < 10 Å (*smaller rectangles*) and <15 Å (*largest rectangles*) and for the SuMo method: *circles* for the ligand defined and *crosses* for ligand undefined

[47–49]. Geometry-based methods have shown that the location of active-site residues can be identified by searching for cavities in the protein structure [50] or by docking small molecules onto the structure [51]. Cavity localization *in silico* has been presented on the basis of the characteristics of the normal (perpendicular to the local surface) created for each portion of surface [52]. A complex analysis of protein interfaces and their characteristics versus highly divergent areas is presented in [53]. Several experimental studies have shown that mutations of residues involved in forming interfaces with other proteins or ligands can be replaced to produce more stable but inactive proteins [54–57]. On this basis, several effective algorithms have been developed [58, 59]. Finally, structural analysis coupled with measures of surface hydrophobicity has been used to identify sites on the surfaces of proteins involved in protein–protein interactions [60, 61].

The Critical Assessment of PRedicted Interaction (CAPRI) is oriented on blind prediction of protein complexes with ligands and protein–protein complex creation [62–66]. The results of competitions are presented on a web site (http://www.capri.ebi.ac.uk) and additionally published in *Proteins Struc Func Gen* (2003) 52 (the whole volume is devoted to this problem). The articles presented there show methods aimed at complex creation based mostly on detailed geometric surface analysis focused on the search for irregularities and cavities [67].

The method presented in this paper is especially dedicated for proteins of active site localized deep in the protein body. The amino acids of high negative values of ΔH_i are understood as an area of higher than expected hydrophobicity. Those localized on the surface may suggest the area of potential protein–protein interaction area. The amino acids of high positive values of ΔH_i are interpreted as the area of lower than expected hydrophobicity. These fragments may be localized in close vicinity of the cavity ready to bind the ligand. The limitation of the described method is under consideration currently and will be published soon.

The results obtained by *fuzzy-oil-drop* model confronted with SuMo results and the data based on the crystal structure seem to be satisfactory. The best agreement is achieved for ribonuclease (1RGE) probably due to the high quality of crystallographic model and the diffraction data (measured by R-value) (R-value 0.109, resolution 1.15 Å). The high score given by SuMo appeared for the highest accordance of results with those based on *fuzzy-oil-drop*. The lowest accordance has been obtained for lysozyme (1LMQ). Probably due to poor determination of the protein structure (R-value 0.165, resolution 1.60 Å) from the diffraction experiment. In summary, one may conclude that the accordance between crystal analysis, SuMo- and *fuzzy-oil-drop*-based results seems to be satisfactory.

The commonly accepted model describing the folding process treats the optimal backbone conformation (peptide bond planes) as the initial step in the approach to protein native structure [68–71]. The second step, expressed by hydrophobic collapse, seems to organize the protein molecule, leading to its native structural form. The *fuzzy-oil-drop* model applied to folding BPTI [72], ribonuclease [73], lysozyme [74], and hemoglobin [75] produced structures close to their native ones. The presence of hydrophobic density irregularities seems to be aim-oriented. A folding process exactly following the path determined by the ideal hydrophobicity distribution could produce a soluble molecule unable to create any complex needed to display the biological activity characteristic for a particular protein molecule. The presence of a ligand or a molecule mimicking the prospective ligand seems to be necessary during folding simulation, ensuring the creation of the cavity of high specificity. The folding of α and β chains of hemoglobin in the absence of heme in a *fuzzy-oil-drop* form external force field led toward structure with a regular hydrophobicity distribution [75]. The CASP5 competition [76] gave precedence to protein structure prediction conditioned by the presence of a ligand (hem), giving additional recognition of the necessity for a ligand to participate in the folding process.

The method presented in this paper addresses the localization of the possible binding site. If the binding cavity is localized correctly, the compatible ligand (its shape and chemical characteristics) can be constructed using classical de novo design methods for ligand construction [77–80].

The conclusion given above may influence the construction of models for protein folding process simulation. The folding process recognized experimentally as multi-step process may be shown in a simplified as follows:

$$U \rightarrow I_1 \rightarrow \ldots \rightarrow I_i \rightarrow \ldots \rightarrow N.$$

The unfolded structural form (U) is transformed into the first intermediate (I_1). The number of intermediates (I_i) is unknown and presumably dependents on a particular protein or group of proteins. The final native structure (N) appears probably as the last one in a sequence of conformational modifications.

The model developed (part of which is presented in this paper) assumes the two-step process with following intermediates:

$$U \rightarrow ES \rightarrow LS \rightarrow N$$

The unfolded state (U) representing the starting event adopts the early-stage conformation (ES), which is determined by backbone conformation presented formerly [81–83] and applied to folding simulation of few proteins [68, 83–85]. The sequence-to-structure (ES) relation generalized

in form of contingency table allowed application to any protein under consideration [69, 86, 87].

The *ES* form changed to late-stage (*LS*), which is mostly driven by hydrophobic interaction between side chains is supposed to approach the native structure (*N*) of the protein.

The *LS* conformation of protein can be created *in silico* according to the *fuzzy-oil-drop* model [72, 74, 88]. The results presented in this paper (as well as other results [73, 75, 89, 90]) suggest that the active presence of ligand during the protein folding process influencing mutually conformation of both molecules may ensure creation of highly specific ligand binding cavity [73, 75].

In consequence, the simulation of protein folding process can be presented as follows:

$$U \rightarrow ES \rightarrow LS \overset{\text{LIGAND}}{\hookrightarrow} N$$

Initial results [72–75, 88–90] encourage large-scale computational experiments, suitable for the grid environment. Further simulations are planned as a part of the "Never Born Proteins" project in the application layer of the EUChinaGRID initiative.

The procedure for folding simulation according to the presented model requires the knowledge of ligand molecule properties (e.g., their electronic structures, partial charges, van der Waals parameters). Presently, such detailed chemical characteristics of ligand (substrate) molecules is not available in a consistent form. Therefore, one of our goals is to create the library delivering the parameters describing the molecules found as protein ligand in the Protein Data Bank (PDB) [91]. The ligand properties are calculated using quantum chemical methods. The optimal geometry of ground state and the atomic partial charges of isolated ligand molecules as well as their energy are calculated using the Amsterdam Density Functional (ADF) [92] in the TZV extended basis set (Slater type functions) with BP86 exchange correlation functional. The atomic charges are calculated according to Mulliken population analysis [93] and Hirshfeld analysis [94]. We plan, also, enclose similar properties computed by Gaussian package [95] for comparison (e.g., partial charges implying from the fitting of electrostatic potential [96]). The service is progressively available on the http://www.bioinformatics.cm-uj.krakow.pl/ligand website.

Availability

The active site can be easily predicted for any protein structure according to the presented model, with the free prediction server available at http://www.bioinformatics.cm-uj.krakow.pl/activesite.

Acknowledgements This research was supported by the Polish State Committee for Scientific Research (KBN), grant 3 T11F 003 28, and Collegium Medicum grants 501/P/133/L and WŁ/222/P/L. The work has been supported by the European Commission EuChinaGRID project (contract number 026634). Chemical quantum calculations for ligand database were performed at the Research Centre in Juelich (project no. 2249) – ADF and at the Academic Computer Centre Cyfronet AGH - Gaussian. Many thanks to Olga Stepien for the implementation of the web interface.

References

1. Kauzmann W (1959) Adv Protein Chem 14:1–63
2. Klapper MH (1971) Biochim Biophys Acta 229:557–566
3. Klotz IM (1970) Arch Biochem Biophys 138:704–706
4. Meirovitch H, Scheraga HA (1980) Macromolecules 13:1406–1414
5. Meirovitch H, Scheraga HA (1980) Macromolecules 13:1398–1405
6. Kyte J, Doolittle RF (1982) J Mol Biol 157:105–132
7. Meirovitch H, Scheraga HA (1981) Macromolecules 13:340–345
8. Rose GD, Roy S (1980) Proc Natl Acad Sci USA 77:4643–4647
9. Baumann G, Frommel C, Sander C (1989) Protein Eng 2:329–334
10. Bonneau R, Strauss CE, Baker D (2001) Proteins 43:1–11
11. Holm H, Sander C (1992) J Mol Biol 225:93–105
12. Novotny J, Rashin AA, Bruccoleri RE (1988) Proteins 4:19–30
13. Irbäck A, Peterson C, Potthast F (1996) Proc Natl Acad Sci USA 93:9533–9538
14. Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W (1982) Faraday Symp Chem Soc 17:109–120
15. Silverman BD (2001) Proc Natl Acad Sci USA 98:4996–5001
16. Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) J Mol Biol 179:125–142
17. Engelman DM, Zaccai G (1980) Proc Natl Acad Sci USA 77:5894–5898
18. Rees DC, DeAntonio L, Eisenberg D (1989) Science 245:510–513
19. Silverman BD (2003) Protein Sci 12:586–599
20. Baldwin RL (2002) Science 295:1657–1658
21. Dill KA (1990) Biochemistry 29:7133–7155
22. Finney JL, Bowron DT, Daniel RM, Timmins PA, Roberts MA (2003) Biophys Chem 105:391–409
23. Crippen GM, Kuntz ID (1978) Int J Pept Protein Res 12:47–56
24. Kuntz ID, Crippen GM (1979) Int J Pept Protein Res 13:223–228
25. Vojtechovsky J, Chu K, Berendzen J, Sweet RM, Schlichting I (1999) Biophys J 77:2153–2174
26. Eschenburg S, Genov N, Peters K, Fittkau S, Stoeva S, Wilson KS, Betzel C (1998) Eur J Biochem 257:309–318
27. Reverter D, Fernandez-Catalan C, Baumgartner R, Pfander R, Huber R, Bode W, Vendrell J, Holak TA, Aviles FX (2000) Nat Struct Biol 7:322–328
28. Neidhart D, Wei Y, Cassidy C, Lin J, Cleland WW, Frey PA (2001) Biochemistry 40:2439–2447
29. Karlsen S, Hough E (1995) Acta Crystallogr D Biol Crystallogr 51:962–978
30. Sevcik J, Dauter Z, Lamzin VS, Wilson KS (1996) Acta Crystallogr D Biol Crystallogr 52:327–344
31. Graille M, Zhou CZ, Receveur-Brechot V, Collinet B, Declerck N, van Tilbeurgh H (2005) J Biol Chem 280:14780–14789
32. Carvalho AL, Dias FM, Prates JA, Nagy T, Gilbert HJ, Davies GJ, Ferreira LM, Romao MJ, Fontes CM (2003) Proc Natl Acad Sci USA 100:13809–13814
33. Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R (2004) Nature 429:780–784
34. Levitt M (1976) J Mol Biol 104:59–107

35. Engelman DM, Steitz TA, Goldman A (1986) Annu Rev Biophys Biophys Chem 15:321–353
36. Hopp TP, Woods KR (1981) Proc Natl Acad Sci USA 78:3824–3828
37. Janin J (1979) Nature 277:491–492
38. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Science 229:834–838
39. Wimley WC, White SH (1996) Nat Struct Biol 3:842–848
40. Wolfenden R, Andersson L, Cullis PM, Southgate CC (1981) Biochemistry 20:849–855
41. Jambon M, Imberty A, Deléage G, Geourjon C (2003) Proteins 52:137–145
42. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S (1999) Nat Genet 23:151–157
43. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) J Mol Biol 283:707–725
44. Skolnick J, Fetrow JS (2000) Trends Biotechnol 18:34–39
45. Wallace AC, Borkakoti N, Thornton JM (1997) Protein Sci 6:2308–2323
46. Zvelebil MJ, Sternberg MJ (1988) Protein Eng 2:127–138
47. Devos D, Valencia A (2000) Proteins 41:98–107
48. Hegyi H, Gerstein M (1999) J Mol Biol 288:147–164
49. Wilson CA, Kreychman J, Gerstein M (2000) J Mol Biol 297:233–249
50. Liang J, Edelsbrunner H, Woodward C (1998) Protein Sci 7:1884–1897
51. Oshiro CM, Kuntz ID, Dixon JS (1995) J Comput Aided Mol Des 9:113–130
52. Lamb ML, Burdick KW, Toba S, Young MM, Skillman AG, Zou X, Arnold JR, Kuntz ID (2001) Proteins 42:296–318
53. Jimenez JL (2005) Proteins 59:757–764
54. Kanaya S, Oobatake M, Liu Y (1996) J Biol Chem 271:32729–32736
55. Meiering EM, Serrano L, Fersht AR (1992) J Mol Biol 225:585–589
56. Shoichet BK, Baase WA, Kuroki R, Matthews BW (1995) Proc Natl Acad Sci USA 92:452–456
57. Zhang J, Liu ZP, Jones TA, Gierasch LM, Sambrook JF (1992) Proteins 13:87–99
58. Elcock AH (2001) J Mol Biol 312:885–896
59. Ondrechen MJ, Clifton JG, Ringe D (2001) Proc Natl Acad Sci USA 98:12473–12478
60. Jones S, Thornton JM (1997) J Mol Biol 272:121–132
61. Jones S, Thornton JM (1997) J Mol Biol 272:133–143
62. Janin J (2005) Proteins 60:170–175
63. Janin J (2005) Protein Sci 14:278–283
64. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ (2003) Proteins 52:2–9
65. Mendez R, Leplae R, De Maria L, Wodak SJ (2003) Proteins 52:51–67
66. Mendez R, Leplae R, Lensink MF, Wodak SJ (2005) Proteins 60:150–169
67. Lei H, Duan Y (2004) Protein Eng Des Sel 17:837–845
68. Brylinski M, Jurkowski W, Konieczny L, Roterman I (2004) Bioinformatics 20:199–205
69. Brylinski M, Konieczny L, Czerwonko P, Jurkowski W, Roterman I (2005) J Biomed Biotechnol 2:65–79
70. Dobson CM (2001) Philos Trans R Soc Lond B Biol Sci 356:133–145
71. Hayward S (2001) Protein Sci 10:2219–2227
72. Brylinski M, Konieczny L, Roterman I (2006) Biochimie 88:1229–1239
73. Brylinski M, Konieczny L, Roterman I (2006) Comp Biol Chem 30:255–267
74. Brylinski M, Konieczny L, Roterman I (2006) J Biomol Struct Dynam 23:519–527
75. Brylinski M, Konieczny L, Roterman I (2006) International Journal of Bioinformatics Research and Applications (IJBRA) (in press)
76. Venclovas C (2003) Proteins 53(Suppl 6):380–388
77. Baurin N, Vangrevelinghe E, Morin-Allory L, Merour JY, Renard P, Payard M, Guillaumet G, Marot C (2000) J Med Chem 43:1109–1122
78. Cramer RD, Patterson DE, Bunce JD (1989) Prog Clin Biol Res 291:161–165
79. Polanski J, Walczak B (2000) Comput Chem 24:615–625
80. Sippl W (2002) J Comput Aided Mol Des 16:825–830
81. Roterman I (1995) J Theor Biol 177:283–288
82. Roterman I (1995) Biochimie 77:204–216
83. Jurkowski W, Brylinski M, Konieczny L, Wisniowski Z, Roterman I (2004) Proteins: Struct Funct, Bioinform 55:115–127
84. Jurkowski W, Brylinski M, Konieczny L, Roterman I (2004) J Biomol Struct Dynam 22:149–157
85. Brylinski M, Jurkowski W, Konieczny L, Roterman I (2004) TASK Quarterly 8:413–422
86. Meus J, Brylinski M, Piwowar M, Piwowar P, Wisniowski Z, Stefaniak J, Konieczny L, Surówka G, Roterman I (2006) Med Sci Monit 12:BR208–BR214
87. Brylinski M, Konieczny L, Roterman I (2004) In Silico Biol 5:0022
88. Konieczny L, Brylinski M, Roterman I (2006) In Silico Biol 6:15–22
89. Brylinski M, Konieczny L, Roterman I (2006) Bioinformation 1:127–129
90. Brylinski M, Kochanczyk M, Konieczny L, Roterman I (2006) In Silico Biol 6:0052
91. http://www.rcsb.org
92. te Velde G, Bickelhaupt FM, van Gisbergen SJA, Fonseca Guerra C, Baerends EJ, Snijders JG, Ziegler T (2001) J Comput Chem 22:931–967; Fonseca Guerra C, Snijders JG, te Velde G, Baerends EJ (1998) Theor Chem Acc 99:391–403; ADF2006.01, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, http://www.scm.com
93. Mulliken RS (1955) J Chem Phys 23:1833–1840,1841–1846
94. Hirshfeld FL (1977) Theo Chim Acta 44:129–138
95. Gaussian 03, Revision C.02, Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene MLiX, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian, Wallingford CT
96. Breneman CM, Wiberg KB (1990) J Comp Chem 11:361–373